DS C29 – Lead Scoring Case Study

- Priyanka Kumari(priyankakgr2006@gmail.com)
- Devarajan Narayanan(devarajan-narayanan@outlook.com)

Content:

- Problem Statement and Dataset Description
- Analysis Approach
- Results and Conclusion

Problem Statement

An Education company named X education sells online courses to industry professionals. The company markets its courses on several websites and search engines. Once these people land on their website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X education gets a lot of leads, its lead conversion rate is very poor. For example, if, they say, they acquire 100 leads in a day, only about 30% of them are converted.

What we need to do:

1) X education wants our help in selecting the most promising leads, i.e. the leads that are more likely to be converted to paying customers
2) The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have lower conversion chance.

DataSet Description

Below dataset is part of this case study:

- **Leads.csv** --> contains all the information captured from the system and manually entered by Sales executives. It has target variable 'Converted' which indicates whether the lead has been converted or not. The model should be trained to predict this variable.

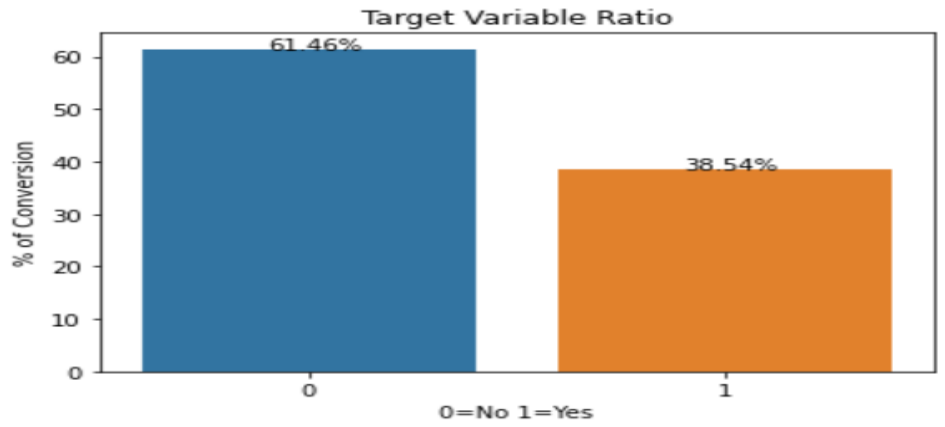
Analysis Approach

1. Began by importing all the necessary libraries like numpy, pandas, sklearn, statsmodel, seaborn, matplotlib etc.
2. Performed the data inspection using below methods
    - Info(), shape, duplicated(), describe(), null checks
3. Performed below data cleaning steps:
    - Replaced all ‘Select’ with np.nan
    - Null Handling
        - Dropped columns having more than 40% null values
        - Performed imputation (replaced null with ‘Others’ and grouped less frequent values to ‘Others’) for below columns:
            - City
            - Specialization
            - Tags
            - What is your current occupation
            - Last Activity
            - Lead Source
            - Last Notable Activity
        - Performed imputation (using median) for below numerical columns:
            - TotalVisits
            - Page Views Per Visit

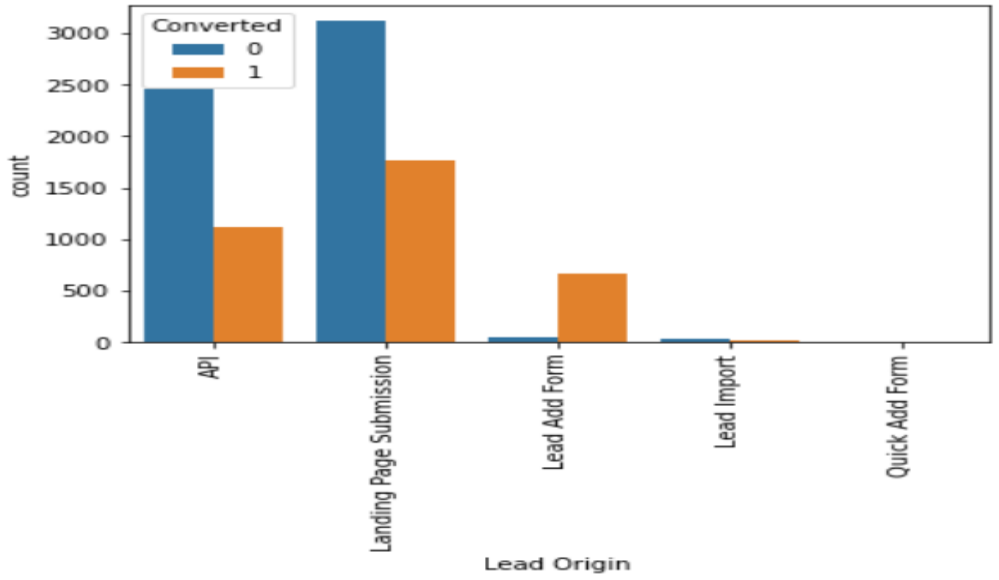
- Further data cleaning steps:
  - Dropped below columns because of highly skewed data distribution:
    - What matters most to you in choosing a course
    - Country
    - Do Not Call/ Do Not Email
    - Search
    - Newspaper Article
    - X Education Forums
    - Newspaper
    - Digital Advertisement
    - Through Recommendation
  - Dropped below columns because of having single category:
    - Magazine
    - Receive More Updates About Our Courses
    - Update me on Supply Chain Content
    - Get updates on DM Content
    - I agree to pay the amount through cheque
  - Outlier Treatment for below columns by imputing the outliers with 1.5*IQR
    - TotalVisits
    - Page Views Per Visit

4. Exploratory Data Analysis:

- Percentage of Converted data (~39%):

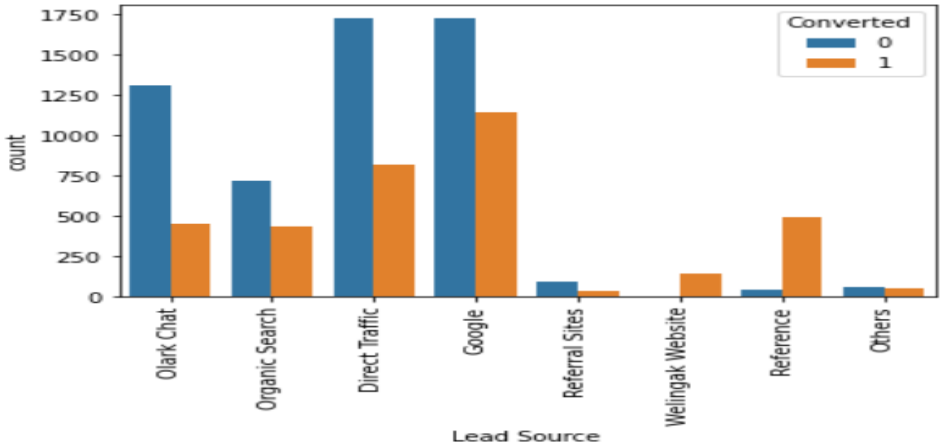


- Lead Origin Vs Converted:



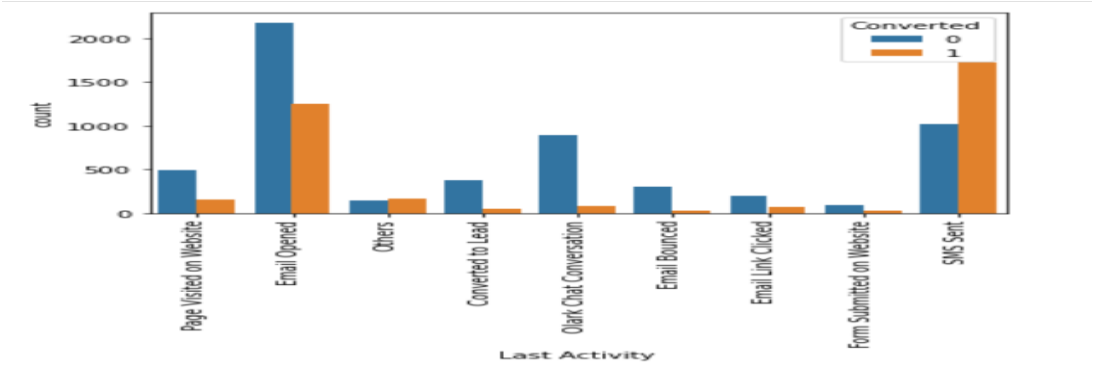
- Inference:
  - API and Landing Page Submission have around 35-40% conversion rate but count of lead originated from them are considerable.
  - Lead Add Form has more than 90% conversion rate but count of lead are not very high.
  - Lead Import and Quick Add From are very less in count.

- Lead Source Vs Converted:



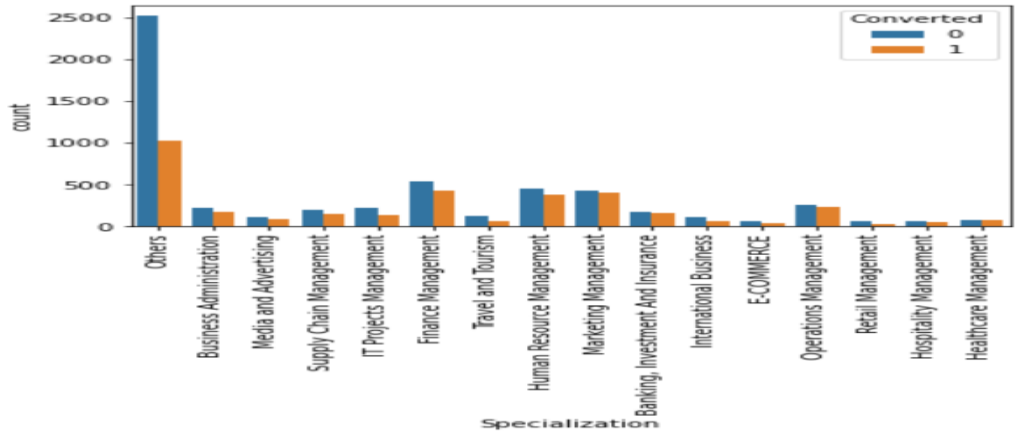
  - Inference:
    - Direct Traffic and Google are generating high number of leads; should be focused for their conversion
    - Welingak and Reference are having more conversion rate hence should be focused to generate more leads
    - leads from Olark Chat and Organic Search should be focused for conversion

- Last Activity Vs Converted:



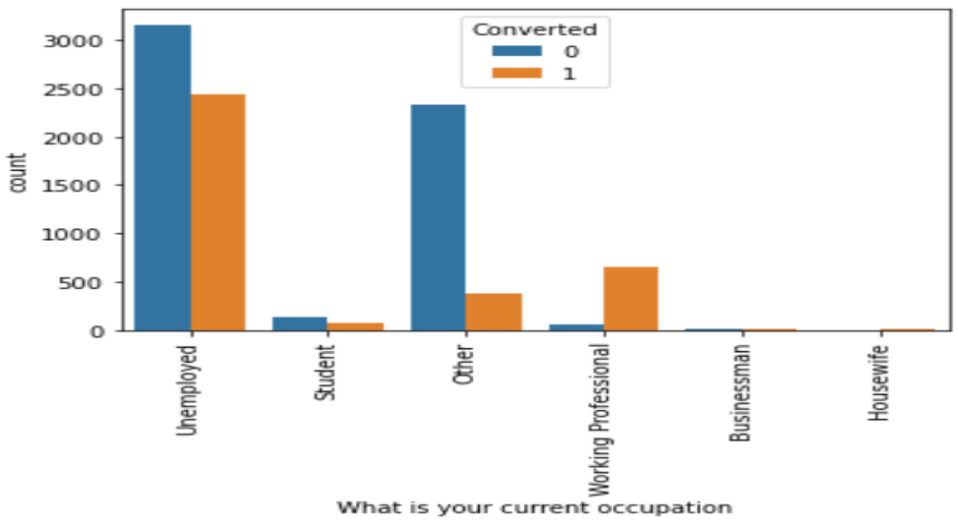
  - Inference:
    - Most of the leads have Email Opened and SMS Sent. More focus should be on leads opening email.
    - conversion rate for SMS sent is very high and hence they should be targeted for lead conversion

- Specialization Vs Converted:



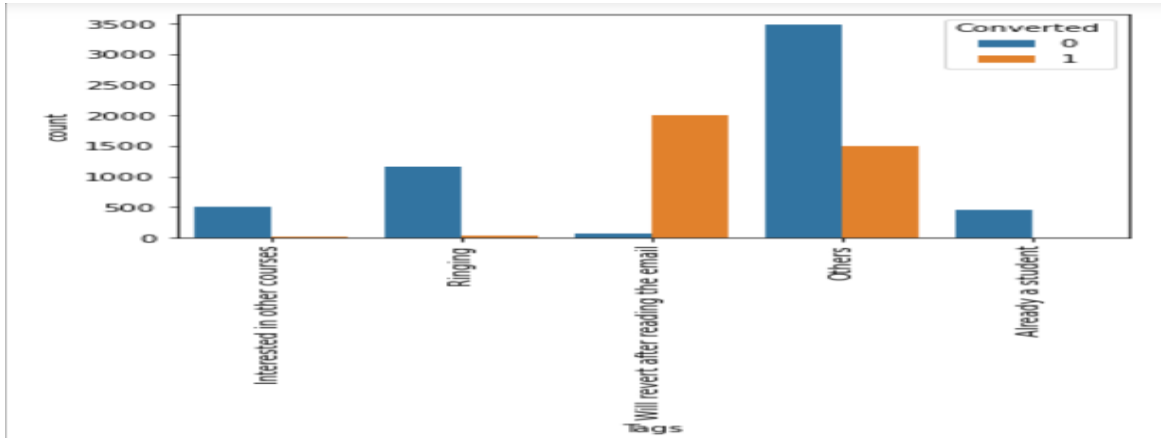
  - Inference:
    - Focus should be on those various specializations with less leads but high conversion rate.
    - Most of the leads are from 'Others'(Not specified) specialization but the conversion rate is poor.

- What is your current occupation Vs Converted:

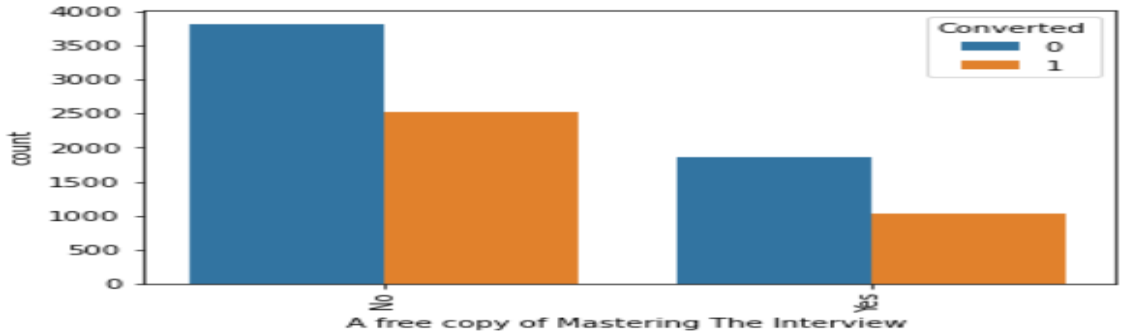


  - Inference:
    - Working professionals have high conversion. focus should be on them for more leads
    - Most of the leads are from Unemployed section, focus should be increasing conversion from them

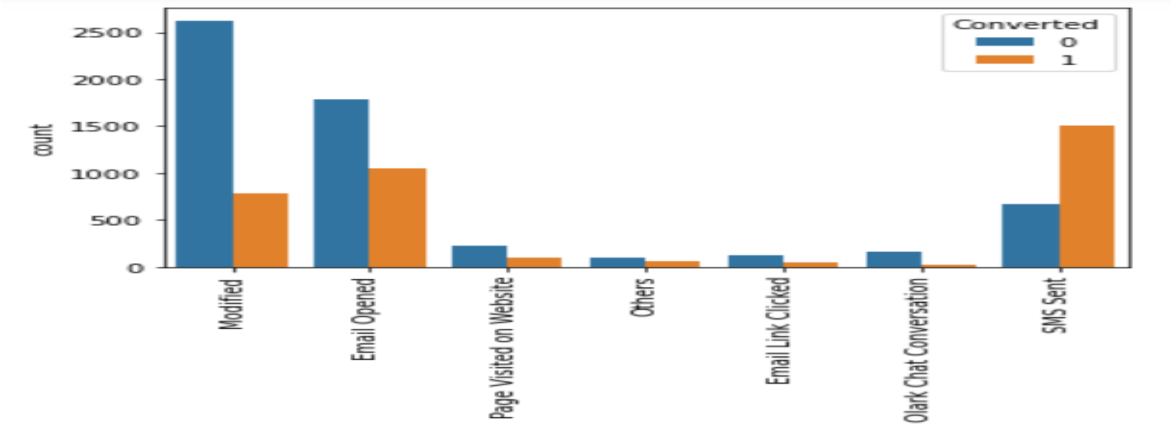
- Tags Vs Converted:



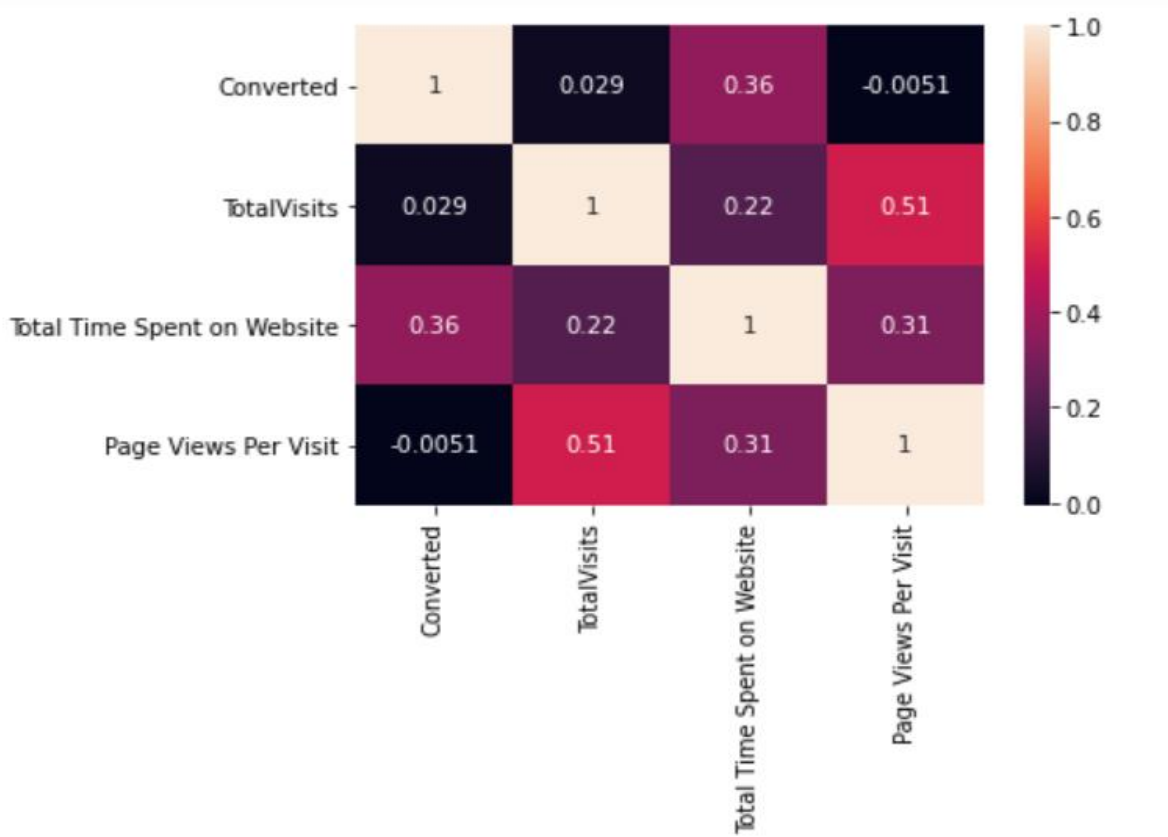
- Inference:
  - 'Will Revert after reading the email' have high conversion. focus should be on them for more leads
  - Most of the leads are from 'Others'(Unspecified) section, focus should be increasing conversion from them
- A free copy of Mastering The Interview Vs Converted:



  - Inference:
    - More leads are from those who do not ask for free copy of Mastering Interviews. Can be focused for conversion
    - There is not much impact of this on the target conversion rate.
- Last Notable Activity Vs Converted:



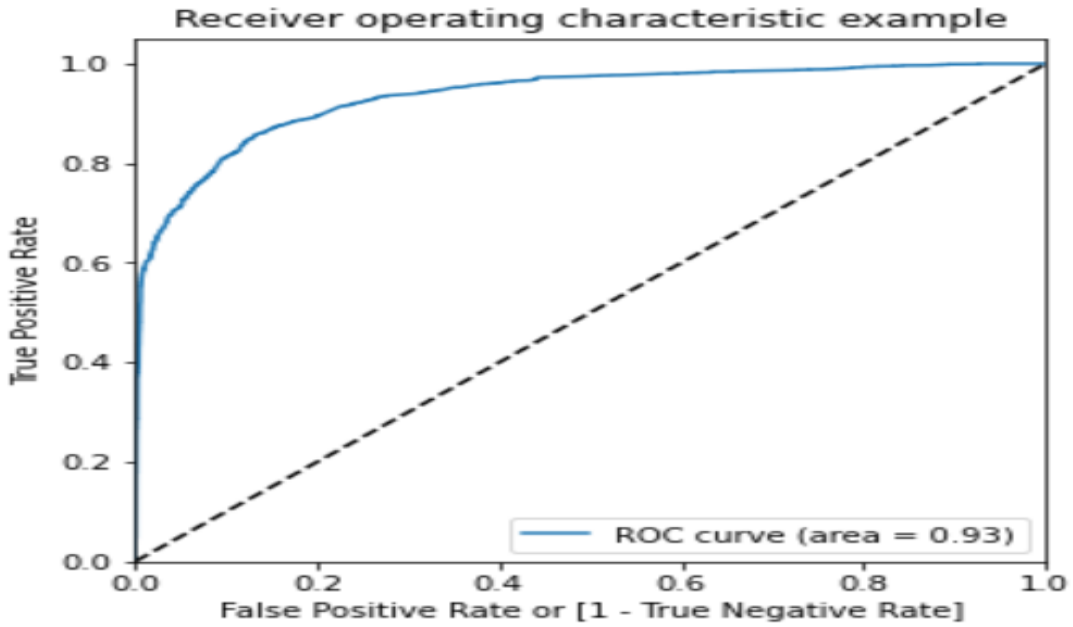
- Inference:
  - More leads are from those who have modified their account or opened email. focus should be on their conversion
  - SMS sent have high conversion
- Numerical Columns (Heatmap to show correlation):



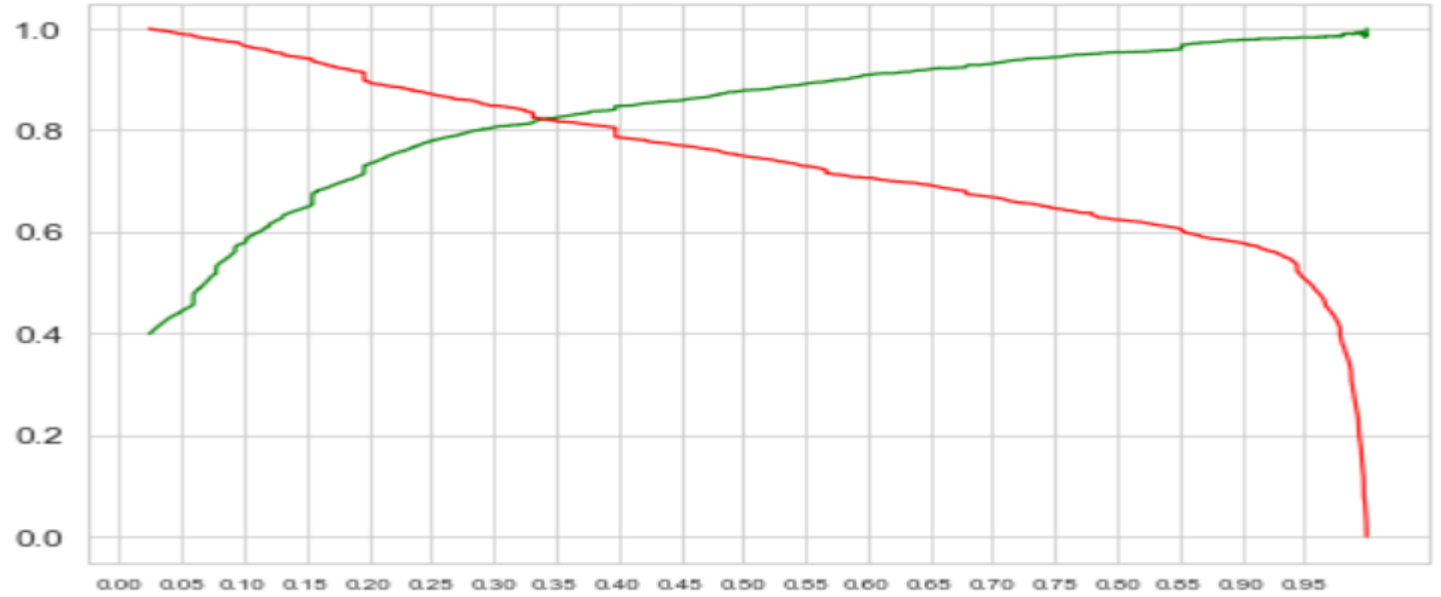
- Inference:
  - Strong correlation between Total Visits and Page Views Per Visits
  - Converted has good correlation with Total time spent on Website

5. Dummy Creation for the variable ‘A free copy of Mastering The Interview’
6. Created dummies for remaining categorical columns
7. Train test Split keeping the 70-30 ratio
8. Scaled the numerical variables ('TotalVisits','Total Time Spent on Website') using StandarScaler
9. Model Building process
    1. Selected top 20 features using RFE method
    2. Built model-1 using statsmodel and verified the summary to check P-value
    3. Dropped below columns due to either P-value and/or high VIF one by one and rebuilt new model(Total 10 iterations):
        - What is your current occupation_Housewife
        - Last Notable Activity_Others
        - Tags_Ringing
        - Specialization_Hospitality Management
        - Lead Source_Reference
        - Last Notable Activity_Email Opened
        - Tags_Others
        - Specialization_Travel and Tourism
        - Specialization_International Business
10. Probability prediction using 10^{th} model on the X_train dataset
11. Created y_train_pred_final dataset which contains original converted indicator and predicted probability

12. By taking an initial cut-off or 0.5, predicted the converted indicator
13. Generated confusion matrix and observed below evaluation matrix:
    - Accuracy score → 86%
    - Sensitivity → 75%
    - Specificity → 93%
    - Precision → 86%
    - Recall → 86%
    - TPR/FPR → 75%, 6%
14. Plotted ROC curve to check the area under the curve to see if the model is good enough or not. Higher the area under the curve, better the model.

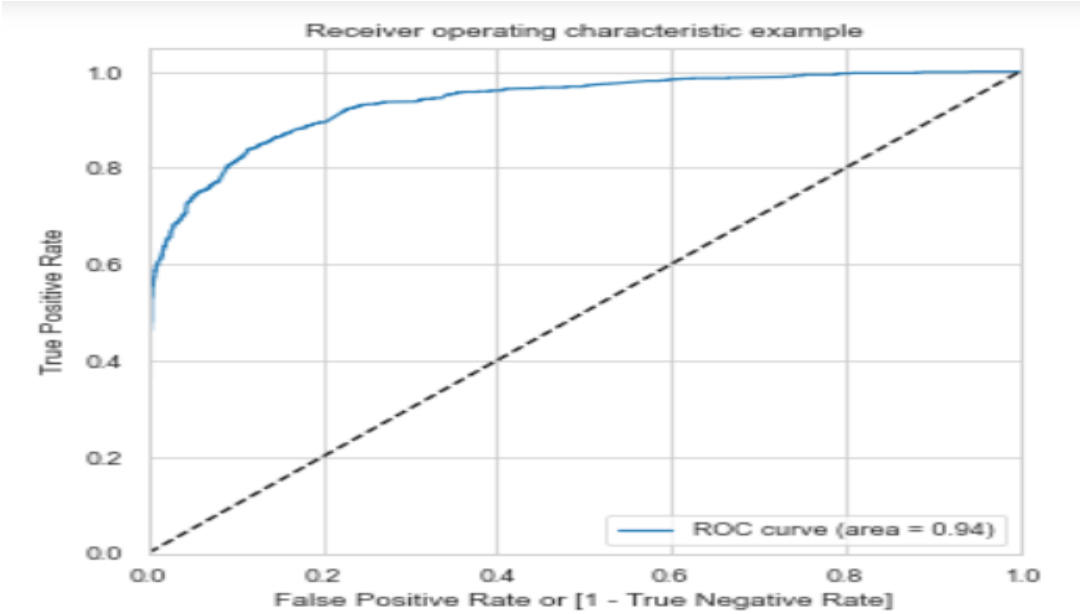


15. Tried to find optimal cut-off point by iterating through 0.0 to 0.9 cut-off and plotted the corresponding accuracy, sensitivity, specificity graph. The new cut-off was found to be 0.28
16. Re -generated confusion matrix and observed below evaluation matrix:
    - Accuracy score → 86%
    - Sensitivity → 86%
    - Specificity → 86%
    - Precision → 80%
    - Recall → 86%
    - TPR/FPR → 86%, 13%
17. Plotted the Precision and Recall trade-off to find new optimal cut-off which came out to be 0.33



18. Re-generated confusion matrix and observed below evaluation matrix:
    - Accuracy score → 86%
    - Sensitivity → 84%
    - Specificity → 88%
    - Precision → 81%
    - Recall → 83%
    - TPR/FPR → 84%, 11%
19. Performed prediction on test data set by first performing scaling and then predicted the conversion probability using the trained model
20. Using the converted probability on test data, evaluated the predicted conversion indicator. 0.33 was taken as threshold here
21. Checked below evaluation matrix:
    - Accuracy score → 86%
    - Sensitivity → 84%
    - Specificity → 88%
    - Precision → 83%
    - Recall → 84%

22. Plotted the ROC curve as shown below:



23. Generated the lead score by 100 to the conversion probabilities.

Features Impacting the target Variable (Converted)

Below shown in the summary report:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6455
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2000.2
Date:	Mon, 09 Aug 2021	**Deviance:**	4000.4
Time:	11:58:32	**Pearson chi2:**	9.10e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3569	0.121	-19.402	0.000	-2.595	-2.119
Total Time Spent on Website	1.0415	0.045	23.131	0.000	0.953	1.130
Lead Origin_Lead Add Form	3.4533	0.212	16.313	0.000	3.038	3.868
Lead Source_Olark Chat	1.2870	0.117	10.970	0.000	1.057	1.517
Lead Source_Welingak Website	2.6065	0.749	3.480	0.001	1.138	4.075
Last Activity_Email Bounced	-1.0352	0.358	-2.889	0.004	-1.738	-0.333
Last Activity_Email Opened	0.5837	0.118	4.957	0.000	0.353	0.815
Last Activity_SMS Sent	1.5791	0.120	13.180	0.000	1.344	1.814
Specialization_International Business	-0.5566	0.328	-1.695	0.090	-1.200	0.087
What is your current occupation_Other	-0.2861	0.093	-3.087	0.002	-0.468	-0.104
What is your current occupation_Working Professional	1.5953	0.239	6.688	0.000	1.128	2.063
Tags_Will revert after reading the email	4.2410	0.170	24.957	0.000	3.908	4.574
Last Notable Activity_Modified	-0.4828	0.098	-4.924	0.000	-0.675	-0.291

Features Impacting the target Variable (Converted)

Sorting features according to their contribution to target conversion

```
log_model_10.params.sort_values(ascending=False)
```

```
Tags_Will revert after reading the email                4.235850
Lead Origin_Lead Add Form                               3.457863
Lead Source_Welingak Website                            2.614611
What is your current occupation_Working Professional    1.600484
Last Activity_SMS Sent                                  1.577706
Lead Source_Olark Chat                                  1.298177
Total Time Spent on Website                             1.041867
Last Activity_Email Opened                              0.581933
What is your current occupation_Other                  -0.285401
Last Notable Activity_Modified                         -0.481195
Last Activity_Email Bounced                            -1.041223
const                                                  -2.368335
dtype: float64
```

Top 3 variables:

- Tags
- Lead Origin
- Lead Source

Top 3 Categorical/Dummy variable:

- Tags_Will revert after reading the email
- Lead Origin_Lead Add Form
- Lead Source_Welingak Website

Conclusion / Recommendation

1. Sales team should take below course of action in order to maximize the conversion rate:
    1. Customers who have been tagged as ‘will revert after reading the email’ should be followed up as they are highly likely to be converted
    2. Someone who is filling the form are more likely to be converted as generally people skip the ads. Only interested people spend time on ads and take any action
    3. Working professionals should be followed up as they would have money to pay, and they need course to learn new things to grow in their career
    4. If someone sends SMS, that means they are interested in taking up the course, hence they should be followed up
    5. Based on time spent on the website, customer should be followed up

2. The Sales team can focus on below points apart from making calls:
    1. Sales team can focus on bringing more leads from reference as they have high conversion rates
    2. Many leads are coming from below specializations, but their conversion rate is not good. Possibly either the courses are not attractive, or the price is high. Sales team can work on providing better facilities to them:
        - Finance Management, Human Resource Management, Marketing Management
    3. So many leads are coming from ‘Unemployed’ customers. Some discounts can be offered to them to make the courses affordable
    4. Team can look for reference with existing students with referral rewards